Variability in Shape as a Function of Size

by

Deborah L. Reichert and Charles E. McCulloch

Biometrics Unit

337 Warren Hall

Cornell University

Ithaca, NY 14853

This paper is BU-1069-MA in the Biometrics Unit Mimeo Series.

SUMMARY

Size and shape, although intuitively simple concepts, are difficult to define mathematically in a biologically meaningful way. Darroch and Mosimann (1985, Biometrika, 72:241-252) have proposed definitions of size and shape that incorporate the concept of geometric similarity, and have developed principal components of shape and total variability in shape based on these definitions. We adopt these definitions of size and shape. Using a related definition of the principal components of shape, we characterize the dependence on the size variable of the principal components of shape and total variability in shape. We derive the minimum amount of variability in shape and the corresponding size variable. The practical implications of these results are explored using data on strongylid eggs; for each of several plausible size variables, we determine the variability in shape and the principal components of shape. The amount of variability in shape and the principal components of shape are seen to be sensitive to the choice of size variable.

Keywords: Morphometrics, multivariate statistics, principal components

1. Introduction

Suppose that a set of measurements have been collected on an organism. For example, these measurements could be the distances between certain easily identified points on a fish. The measurements will contain information about both the size and the shape of the organism and also about the relationship between size and shape in a population of organisms.

Defining size as a scalar variable and shape as a vector of dimensionless variables, as proposed by Darroch and Mosimann (1985), we develop principal components of shape and total variability in shape. The amount of variability in shape depends on the specific size variable chosen. We explore the nature of this relationship. In particular, we derive the size variable that produces the least variability in shape, and show that it corresponds to Darroch and Mosimann's "invariant to size" solution.

The proposed approach differs from the approach of Burnaby (1966) and others who seek to separate size and shape into independent quantities. Size and shape in this context, as in real life, are not statistically independent. Instead, shape as we define it here is motivated by geometric similarity. Darroch and Mosimann

(1985) propose comparing the principal components of log measurement and log shape to gain information on the relative roles of size and shape in describing the variability of the data.

2. Definitions and Notation

Let $x' = (x_1, \dots, x_p)$ be a vector of p measurements on an organism. Define a size variable to be a generalized geometric mean represented by

$$g_a(x) = \prod_1^p x_i^{a_i},$$

where $a' = (a_1, \dots, a_p)$ is subject to $a'1 = 1$ and $1' = (1, \dots, 1)$. Shape is a vector derived from this size variable by

$$u_a = \frac{x}{\prod_1^p x_i^{a_i}}.$$

We analyze the logarithms of these quantities, log measurement ($y = \log x$), log size ($a'y = \log g_a(x)$), and log shape ($\log u_a = y - (a'y)1 = (I - 1a')y$). In particular, we analyze the principal components of the sample variance-covariance matrices of log measurement and log shape. While our definitions of size and shape are the same as those used by Darroch and Mosimann (1985), their formulation of the principal components of shape assumes the use of the standard geometric mean as the size variable. We consider the principal components of log shape in a setting that allows the scientist to specify any size variable of the type defined above.

3. Principal Components and Total Variability

Let us first define the principal components and total variability of log measurement. To facilitate the analysis of log shape, principal components of log shape should be comparable to those of log measurement, so we need notation for the components of log measurement that can be extended to cover log shape. Let S denote the sample variance-covariance matrix of log measurement. To define principal components of log measurement, let the first component, $\alpha' y$, have the maximum variance, $\alpha' S \alpha$, over α subject to $\alpha'\alpha = 1$. Denote this component by $\alpha_1' y$. Define the second component, $\alpha_2' y$, as the maximum of $\alpha' S \alpha$ over α subject to $\alpha'\alpha = 1$ and $\alpha_1'\alpha = 0$. Next, define α_3 as above subject to $\alpha_1'\alpha = 0$ and $\alpha_2'\alpha = 0$, and so on.

It is well known that the vectors α_i defined as above satisfy $(S - \lambda_i I)\alpha_i = 0$, where $\lambda_1 \geq \dots \geq \lambda_q > 0$ are the nonzero roots of $|S - \lambda I| = 0$. Moreover, $\sum \lambda_i = \mathrm{tr}(S)$. Let $\sum \lambda_i$, or $\mathrm{tr}(S)$, denote the total variability present in log measurement.

The principal components and total variability of log shape can be derived similarly to those of log measurement. Because log shape is a function of log measurement, specifically, $\log(u_a) = (I - 1a')y$, the variance-covariance matrix of log shape is a function of that of log measurement. Let $T_a = I - 1a'$. Then, as $\text{var}(T_a y) = T_a S T_a'$, $T_a S T_a'$ is the variance-covariance matrix of log shape. We can also find principal components, $\beta_i'(T_a y)$, that maximize $\beta' T_a S T_a' \beta$ subject to $\beta'\beta = 1$ and $\beta_i'\beta_j = 0$ for $i \neq j$ and satisfy $(T_a S T_a' - \mu_i I)\beta_i = 0$. Here $\mu_1 \geq \ldots \geq \mu_s > 0$ are the s nonzero roots of $|T_a S T_a' - \mu I| = 0$. Furthermore, $\sum\mu_i = \text{tr}(T_a S T_a')$. The total variability in log shape is defined to be $\sum\mu_i$, or $\text{tr}(T_a S T_a')$.

4. Relationship Between Variability in Measurement and in Shape

Total variability in log measurement is related to total variability in log shape as follows:

$$\sum_1^q \lambda_i = \sum_1^s \mu_i - pa'Sa + 21'Sa. \tag{4.1}$$

If we consider $1a'y$, the vector corresponding to the size variable $a'y$, we see that its variance is $pa'Sa$ and that $\text{cov}(1a'y, T_a y) = T_a S a' 1$. Since we cannot, in general, reduce (4.1) to terms involving only $a'Sa$ and $\sum\mu_i$, we cannot partition the variability in log measurement into variability due to log size and variability attributed to log shape. The correlation between size and shape reflects the situation found in the growth patterns of most species, that organisms of different sizes commonly have different shapes as well. However, it renders risky any attempt to distinguish the relative roles of size and shape in contributing to the variability in the data.

For a given sample, $\mathrm{tr}(S) = \sum_1^q \lambda_i$ is a constant. Therefore, looking back at (4.1), we see that total variability in log shape as a function of log size is a convex elliptic paraboloid. As such, there is no maximum amount of variability in log shape. In fact, variability in log shape can exceed variability in log measurement if $2\mathbf{1}'S\boldsymbol{a}$ is a sufficiently large negative number. Indeed, variability in log shape is unbounded and can be made arbitrarily large by appropriate choice of size variable, or $\boldsymbol{a}$.

5. Minimum Variability in Log Shape

The minimum over all size variables of total variability in log shape provides a lower bound on the variability that must be attributed to shape. The minimization of $\sum \mu_i$ as in (4.1) over $\boldsymbol{a}$ subject to $\boldsymbol{a}'\mathbf{1} = 1$ is straightforward. When S^{-1} exists, as it almost always will for continuous data, the value of a that minimizes $\sum\mu_i$ is $a^* = p^{-1}1$. In this case, $a^{*\prime}y$ represents the log of the geometric mean of the data. Because S is positive definite, a^* does indeed define a minimum amount of variability in log shape. For a^*, the amount of variability attributed to shape is

$$\sum_1^s\mu_i = \sum_1^q\lambda_i - p^{-1}1'S1. \tag{5.1}$$

Thus, the size variable that produces the minimum amount of variability in log shape is independent of S, the variance-covariance matrix of the measurements. Only the amount of variability in log shape for this size variable depends on S. Although this solution follows directly from the original formulation of the problem, it produces an intuitively uncomfortable situation. Intuitively, it seems that the size variable that minimizes the variability in log shape should depend on the variance-covariance structure of the measurements.

Should S be singular, which will happen only rarely, if ever, $a^* = p^{-1}1$ will still minimize $\sum\mu_i$, and (5.1) will hold for any a, including a^*, that minimizes $\sum\mu_i$.

6. Example

We illustrate the practical repercussions of these results by considering the following data, which are measurements on the eggs of an equine intestinal parasite, Strongylus vulgaris. The data were collected by Jay Georgi of the New York State College of Veterinary Medicine. The four measurements are as follows: greatest diameter of the egg (length), greatest diameter perpendicular to length (width), and the square roots of the areas at the poles of the egg. Each pole area is defined as the area subtended by the chord perpendicular to the length at a distance of one-twentieth of the length from the end of the egg. There are data for 100 eggs.

We consider three different size variables for this data, each of which might be appropriate to certain specific applications. These size variables are the length of the egg, the crossectional area of the egg (the product of length and width), and the standard geometric mean of the four measurements. To comply with the condition that the size variable be a generalized geometric mean, we use the square root of the crossectional area of the egg.

The first principal components of log shape and the total variability in log shape for each of these size variables is presented in table 1. To facilitate comparison of principal components between measurement and shape and between shapes defined by different size variables, the coefficients of the principal components have been transformed to apply to log length, log width, log pole area 1, and log pole area 2. Without the transformation, the coefficients of the principal components apply to the log shape variables, e.g., if length is the size variable, to log width – log length, log pole area 1 – log length, and so on. The principal components and transformed principal components are completely equivalent.

This example illustrates that the results of a principal component analysis of log shape are heavily dependent on the choice of size variable. When length is the size variable, the first principal component indicates that longer eggs are narrower, whereas for the other size variables, the first principal component indicates that longer eggs are wider. These differences are substantial enough to imply different interpretations of the underlying structure of the data, yet they all derive originally from the same data. The amount of variability in log shape as a proportion of the variability in log measurement also changes in response to the choices of size variables, ranging in these cases from 114 percent to 63 percent of the variability in log measurement. Clearly, the choice of a size variable must be considered carefully before a principal component analysis of shape is performed.

7. Conclusions

While biologists might like to partition the variability in their data into variability due to size and variability due to shape, or at least to assign relative roles to size and shape, no such partitioning is possible with these definitions of size and shape. Darroch and Mosimann (1985) propose using the ratio of total variability in log shape to total variability in log measurement as a measure of the importance of shape in the data, but the unbounded nature of the ratio under different size variables renders it unhelpful. Some degree of comparability at least may be achieved at the minimum value of the ratio, but even then, this quantity lacks biological meaning because of the covariance of size and shape. The desired comparability also suffers from the dependence on the data of this covariance.

Both variability in log shape and the principal components of log shape vary with the choice of size variable that defines shape. The principal components can change enough from one size variable to another that their biological interpretations change. Care must therefore be taken in selecting a size variable when a principal components analysis of shape is planned.

Further study of how the principal components of log shape, especially the first one, change as a function of size would help to further evaluate the sensitivity of principal components analyses of shape to the definition of the size variable. However, the amount of difference between sets of principal components will be hard to quantify in a biologically interpretable way.

All of the foregoing assumes that shape can be effectively defined as a unitless, or dimension-free, quantity. This will not always be possible, however. For example, an analysis of leaves could not readily incorporate information about number of lobes in this framework, although such information is arguably shape information. For the present, at least, morphometricians need to continue to explore and develop new conceptions of size and shape and analyses appropriate for them.

Table 1

First Principal Components and Total Variability in Log Shape for Different Choices of Size

		First Principal Components				total variability in
		Variable				log measurement
		length	width	pole 1	pole 2	
log measurement	$\alpha_1 =$	(0.11	0.01	0.88	0.47)	17.1
log size						total variability in
						log shape
length	$\beta_1 =$	(-0.85	0.16	0.40	0.29)	19.6
crossect. area	$\beta_1 =$	(-0.47	-0.51	0.62	0.36)	15.5
geometric mean	$\beta_1 =$	(-0.37	-0.48	0.79	0.06)	10.7

REFERENCES

Burnaby, T. P. (1966). Growth-Invariant Discriminant Functions and Generalized Distances. Biometrics 22, 96-110.

Darroch, J. N., and Mosimann, J. E. (1985). Canonical and Principal Components of Shape. Biometrika 72, 241-252.